Paper SAS6223-2016

What's New in SAS® Federation Server 4.2

Tatyana Petrova, SAS Institute Inc., Cary NC

ABSTRACT

The SAS® Federation Server is a scalable, threaded, multi-user data access server that provides seamless, integrated data from various data sources. When your data becomes too large to move or copy and too sensitive to allow direct access, the powerful set of data virtualization capabilities allows you to effectively and efficiently manage and manipulate data from many sources, without moving or copying the data. This agile data integration framework will allow business users the ability to connect to more data, reduce risks, respond faster, and make better business decisions. For technical users, the framework provides central data control and security, reduces complexity, optimizes commodity hardware, promotes rapid prototyping, and increases staff productivity.

This paper provides an overview of the latest features of the product and includes examples for leveraging product capabilities.

INTRODUCTION

When positioned between analytics applications and underlying data sources, SAS Federation Server enables you to create a centralized abstract layer for governed access to heterogeneous data. It provides support for additional levels of security and centralizes security management and data governance across the logical data warehouse. The SAS Federation Server environment incorporates logic for threaded, multi-user processing of incoming data requests as well as the ability to optimize this processing to minimize data movement between your data warehouse environments.

More details about the value of data federation as well as the business application of SAS Federation Server can be found in “Exploring Data Access Control Strategies for Securing and Strengthening Your Data Assets Using SAS® Federation Server”. (See the References section for more information.) We highly encourage you read that paper for a good business and technical overview of the product.

This paper focuses on the capabilities of SAS Federation Server 4.2 (the latest release). This release runs on the third maintenance release for SAS 9.4.

When talking about SAS Federation Server, it is important to differentiate between the types of users. Here are the typical user personas who would be working in such an environment:

- SAS Federation Server Administrator
- Data Administrator
- Data Architect
- Report Users

All these users can be set up to work with SAS Federation Server either programmatically or by using the SAS Federation Server Manager web application. On the other hand, Report Users are typically leveraging the SAS Federation Server infrastructure implicitly by using the business intelligence or analytics applications that sit on top of this infrastructure. For this audience, the logical data layer is carefully prepared and secured by other user roles.

This paper focuses on enabling SAS Federation Server Administrators, Data Administrators, and Data Architects to understand and use the new 4.2 capabilities to enhance their logical data warehousing environments.

WHAT IS NEW IN SAS FEDERATION SERVER 4.2?

Here is a brief list of the major additions to the 4.2 release. More information about some of these capabilities are discussed later in the paper.

- Integration with SAS Metadata Server and Web Infrastructure Platform (WIP)

  This integration is so important that we dedicated an entire section of this paper to it!

- Support for SAS DS2 – DS2 is a modular DATA Step programming language that enables advanced data manipulations and scalable processing.

  With this release, you can define and execute DS2 code in SAS Federation Server. For more information about the DS2 language, see *SAS 9.4 DS2 Language Reference*.

- Support for embedded Data Quality and Cleansing operations

  SAS Data Quality and Cleansing technologies are well known and widely adopted in the market. This release of SAS Federation Server provides the capability to embed SAS data quality operations within code created and executed from SAS Federation Server. This functionality is especially beneficial when integrating data from multiple data environments where data might be stored using different standards and must be mapped for a unified reporting purpose.

- Push down processing (DS2 code and Data Quality functions) to run in database

  This release of SAS Federation Server is integrated with SAS in-database technologies, such as SAS In-Database Code Accelerator and SAS Data Quality Accelerator (if these products are licensed at the site). This integration provides corresponding data platforms as a run-time environment for your specially-formatted custom DS2 jobs and Data Quality functions that are part of a DS2 code stream. This powerful capability allows you to leverage your investments in large data processing platforms. Learn more about this optimization later in the paper.

- Enhanced dynamic masking

  Dynamic masking is used to hide sensitive information from the data consumer. This capability is well suited for scenarios when data is expected to be present for analysis purposes and with the help of flexible masking rules, gets encrypted in the results presented to consumer.

  This release of SAS Federation Server extended the set of dynamic rules for data masking by adding the TRANC, RANDIG, RANDATE, and RANSTR algorithms.

- Enhancements for in-memory caching

  Caching enhancements include the ability to cache views in the SAS Federation Server Memory Data Store (MDS) as well as the ability to rematerialize cache in memory when you restart the server.

- Capability to share data sources across multiple SAS Federation Servers

  The new driver for SAS Federation Server allows you to establish a connection from one SAS Federation Server to another and federate data between them.

- Access to secured SAS data sets with metadata-bound libraries

  Metadata-bound libraries are Base SAS libraries where the SAS data sets are bound with their metadata registration. No programmatic code can access these data sets without metadata authorization applied.

- The family of data drivers is extended with the new member - native driver for Apache Hive

With the native driver for Apache Hive, you can now query and manage large data sets that reside in distributed Hadoop storage. This driver supports access to multiple Hadoop distributions.

Note that access to another Hadoop family member, Impala, is supported via SAS Federation Server driver for ODBC which is bundled with a corresponding DataDirect ODBC driver.

- SAS Federation Server 4.2 is packaged with SAS Studio

  SAS Studio is an easy-to-use web-based SAS programming client application with drag-and-drop capabilities and predefined tasks. SAS Studio can be used with the LIBNAME engine for SAS Federation Server, so data consumers can work directly with their SAS Federation Server data objects.

- The interface of the SAS Federation Server Manager got a makeover (Figure 1). While its functionality and navigation follow the principles of the previous release, you will see some differences in the appearance. It is possible to select different themes for this appearance. These themes are configurable by using global preferences in SAS Federation Server Manager. SAS Theme Designer for Flex is a separate SAS product that allows creation and deployment of custom themes.

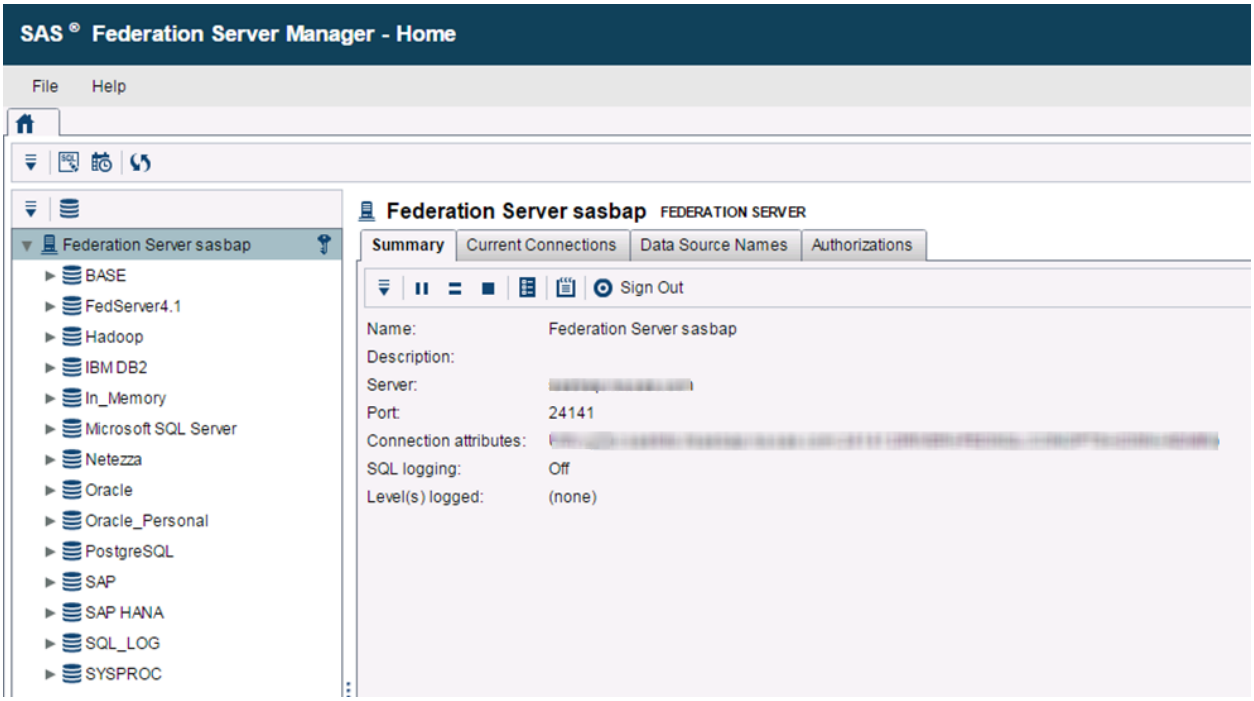

Figure 1. SAS Federation Server Manager 4.2

WHAT IS THE SIGNIFICANCE OF SAS METADATA SERVER INTEGRATION?

In the 4.2 release, SAS Federation Server integrated with SAS® Metadata Server and SAS® Web Infrastructure Platform. For customers familiar with the 4.1 release, note that DataFlux Authentication Server is no longer used by SAS Federation Server except to migrate from 4.1 to 4.2.

This architectural change is significant and has implications on how you approach security design for your SAS Federation Server environment. Figure 2 shows an updated conceptual architecture for the 4.2 release.

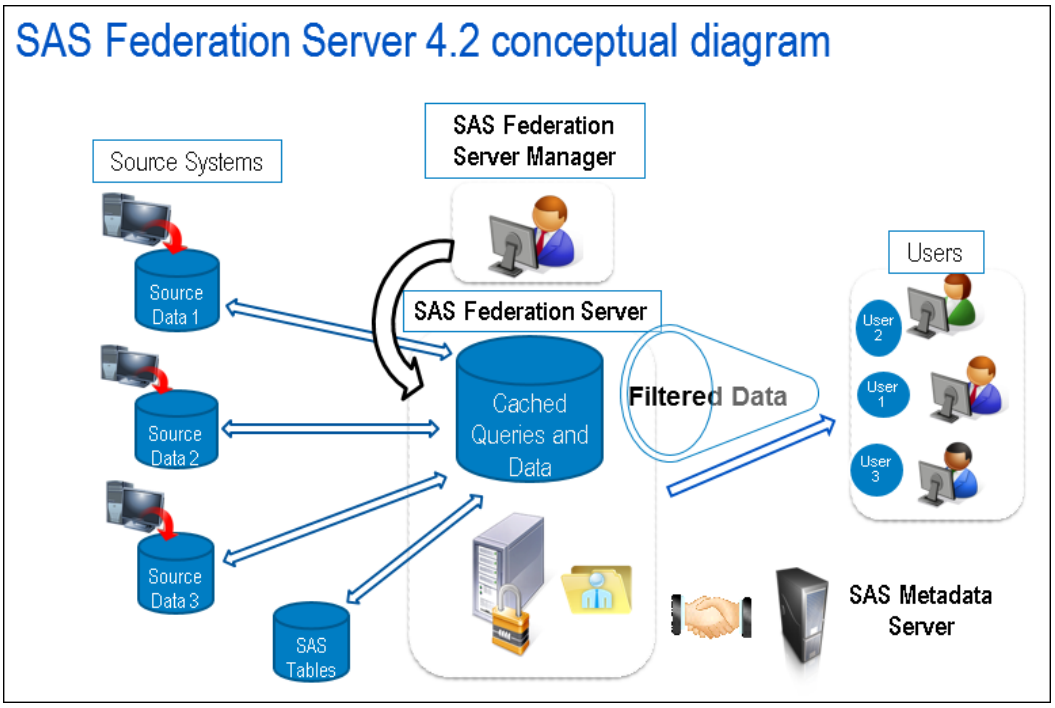

Figure 2. SAS Federation Server 4.2 Conceptual Diagram. Note the Presence of SAS Metadata Server

SAS Metadata Server is used across SAS environments as a central location for storing and managing a wide variety of administration metadata. It also serves as an authentication provider. SAS® Management Console is the standard interface for SAS Metadata Server administration and metadata management. Subsets of this metadata are leveraged and surfaced in many SAS products.

SAS Federation Server 4.2 adopted SAS Metadata Server as an authentication mechanism and a centralized repository for storing and managing user and group configurations.

The following metadata objects used by SAS Federation Server are stored in SAS Metadata Server and configured via SAS Management Console:

- users and user groups
- logins (personal, group, and shared)
- authentication domains

SAS Federation Server provides an additional level of granularity and control for security across your logical data warehouse. Additional types of permissions are configured on the SAS Federation Server side and stored with the SAS Federation Server metadata in its system database (SYSCAT.tdb).

SAS Federation Server also provides a layer of logical data warehouse definitions. Metadata about data platform connections, data services, and underlying dependencies for data hosting objects is also stored in the SAS Federation Server system database.

Figure 3 highlights the hierarchy of metadata objects in SAS Federation Server.

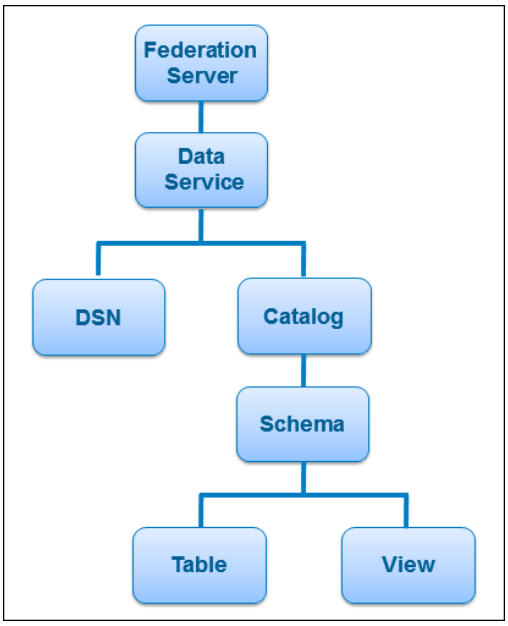

Figure 3. Hierarchy of Metadata Objects in SAS Federation Server

Here is a brief overview of a purpose of these metadata objects:

- **Data Service** contains connection information specific to the driver and data source location. Data services are associated with the authentication domains that are defined in SAS Metadata Server. These services retrieve the credentials for the connecting account.
- **DSN** stands for Data Source Name and goes hand-in-hand with the data service and stores data service' connection parameters. When creating a data service, the DSN is automatically created and adopts the name assigned to a new data service. In addition, the DSN defines how security is enforced for objects parented by a corresponding data service.
- **Catalogs** and **Schemas** are concepts from the grouping of the database objects. Most databases have a concept of a schema that groups tables and views. Many databases also have a concept of a catalog that groups schemas. To uniquely identify each data source across the heterogeneous environment, SAS Federation Server adopted the catalogs and schemas grouping of resources that map to the corresponding grouping on each database side. For databases that do not carry a concept of a catalog or even a schema (for example, for the Base SAS data service), data architects who link resources from this database to SAS Federation Server are prompted to provide logical identifiers for these levels. This prompting provides consistency across the logical data warehouse and allows you to specify unique three-level identifiers for each data source (CATALOG.SCHEMA.TABLE-NAME). You can then reference this three-level name throughout the logical environment.

Data services, catalogs, schemas, tables, and views can have security privileges controlled via SAS Federation Server. The granularity of the privileges goes all the way down to row-level access controls of the underlying source tables. Here are some examples of privileges:

- SELECT (where the SELECT privilege can have a WHERE predicate which effectively implements row-level restrictions)
- UPDATE
- INSERT
- REFERENCES
- DELETE

- EXECUTE
- CREATE, DROP, ALTER TABLE or VIEW
- CREATE or ALTER CACHE
- ADMINISTER
- CONNECT

To summarize, SAS Metadata Server is the place for managing authentication domains, users, groups, and their logins, as well as permissions on SAS Federation Server object itself. SAS Federation Server hosts the lower-level privileges controlled by SAS Federation Server.

In addition, note that security restrictions in the data platforms take precedence over permissions granted by SAS Federation Server. In this way, you have a logical and comprehensive coverage of all the security layers across your enterprise environment.

WHERE ARE MY QUERIES BEING PROCESSED?

One of the benefits of adding SAS Federation Server as a layer between your analytics applications and data warehouses is its intelligence and flexibility around selecting a query execution environment.

Querying performance optimization is based on minimizing data movement, data pre-aggregation, finding the optimal path through data sources, and leveraging the power of distributed execution.

SAS Federation Server brings together multiple SAS data access and in-database technologies, as well as SAS programming language advancements to accommodate for these optimization principles. SAS Federation Server incorporates caching mechanisms as well as automated optimization of the execution path based on the nature of a query and the SAS components licensed.

With SAS Federation Server 4.2, a request for data and data manipulations might be in either FedSQL or DS2 dialects. For both types of requests, data processing can either take place in SAS Federation Server or to a certain degree, in the database. In addition, cached (materialized) views can be created and stored as SAS data sets on disc, in SAS Federation Server memory (Memory Data Store, or simply MDS), or as relational tables in database. All these choices provide optimization opportunities for data architects, database administrators (DBAs), and SAS Federation Server administrators.

With such flexibility comes the importance of understanding of the components and logic used for selecting a run-time environment and corresponding execution flows.

Figures 4 and 5 show several typical execution scenarios based on a type of a user request and environment configuration.

Figure 4 focuses on FedSQL requests, while Figure 5 explains the logic for DS2 code execution.

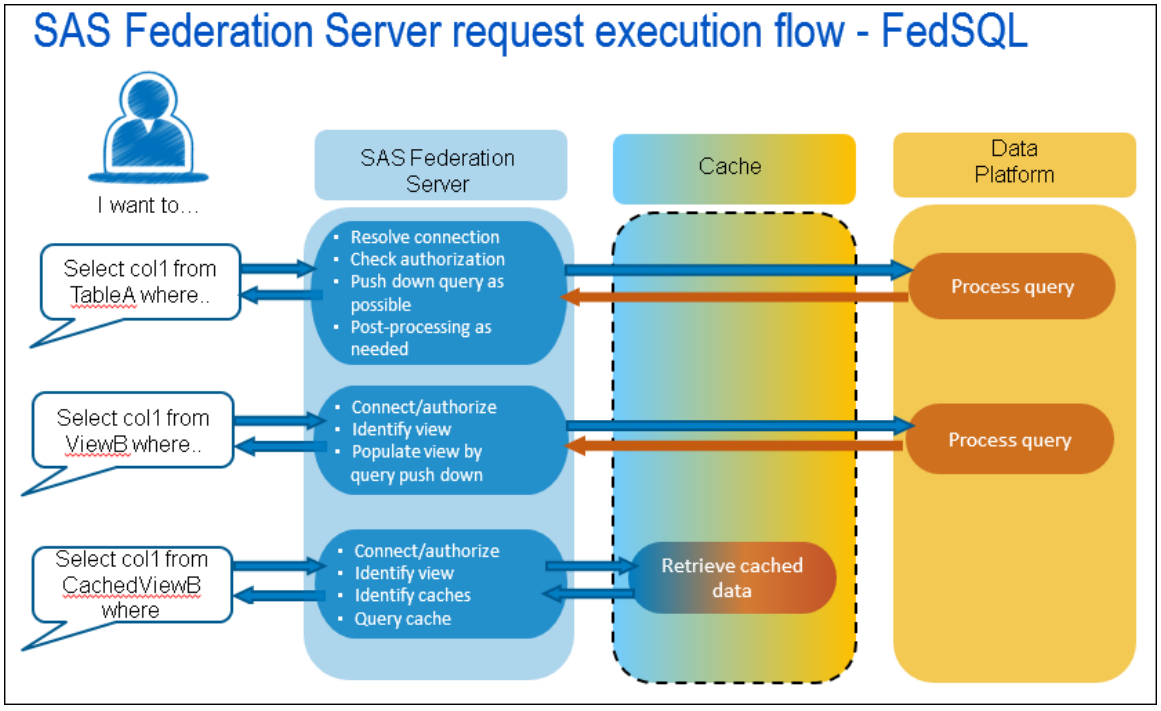

Figure 4. SAS Federation Server Request Execution Flow for a FedSQL Type of Request

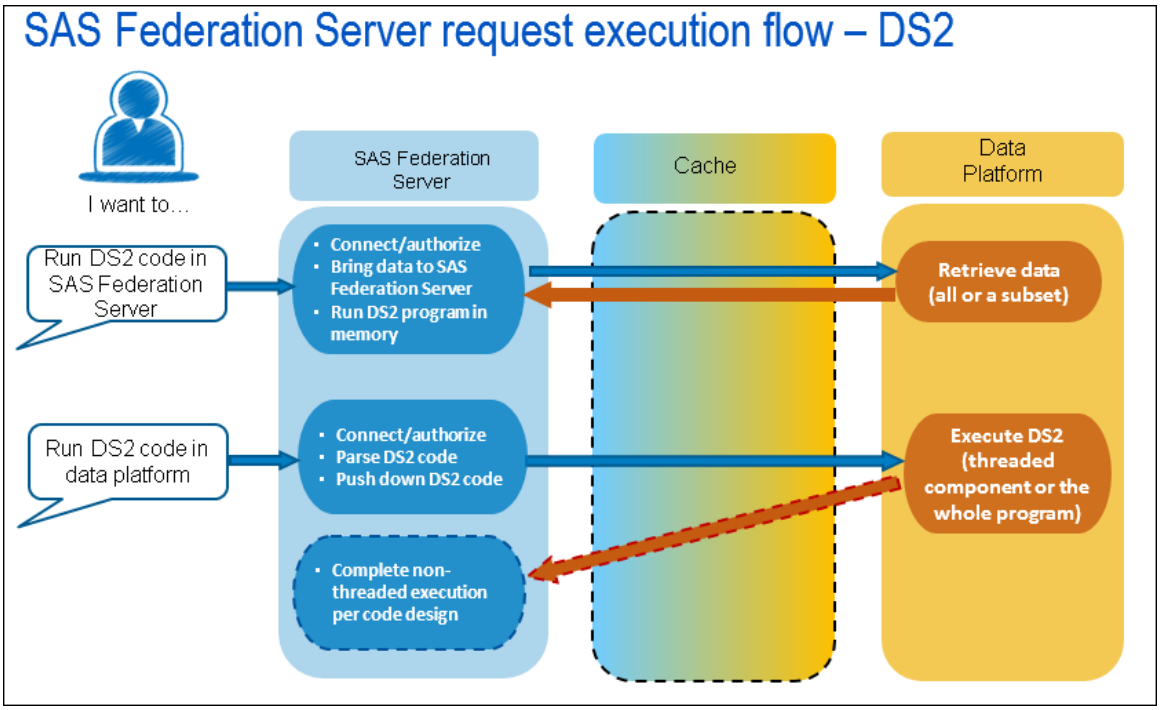

Figure 5. SAS Federation Server Request Execution Flow for a DS2 Type of Request

SAS DS2 code is comprised of threaded and DATA Step components. DS2 code can run fully in SAS Federation Server run-time environment on either cached data or by pulling data from data sources. DS2 program can also use a SQL subquery as an input source, in which case a subquery would be pushed

down to database eliminating the need to bring all the data to SAS Federation Server for a DS2 job execution.

Alternatively, if SAS In-Database Code Accelerator is licensed and configured for the particular data platform, the threaded component of a DS2 program will be pushed down to a database for parallel execution. In the third maintenance release for SAS 9.4, SAS In-Database Code Accelerator is available for Teradata, Hadoop, and Pivotal Greenplum. For Teradata and Hadoop, if DS2 program complies with a set of push down principles, the whole DS2 program would be pushed down for execution keeping data at its source. For more details on DS2 push down principles, see *SAS 9.4 In-Database Products: User's Guide*.

The scenarios listed are not exhaustive. There are variations, especially in cached views utilization for both FedSQL and DS2 scenarios. But similar principles apply.

When evaluating the request, SAS Federation Server weighs the type of the request, location of data sources, resolves the views, identifies if data is cached and where the cache is located, and determines if portions of the request can be pushed down to databases.

For a SAS administrator or DBA conducting a performance evaluation of the data access component of the system, it is critical to validate if the system is configured correctly, so a query issued by SAS can be pushed down to databases when expected to do so. A careful analysis of most typical requests can direct choices for caching certain views for even faster performance.

HOW CAN I IMPROVE THE QUALITY OF MY DATA FROM SAS FEDERATION SERVER 4.2?

SAS Federation Server 4.2 added support for embedded data quality (DQ) functions.

A package of DQ functions (which are standardized with the DQ functions from SAS data management family of products) are implemented using SAS Quality Knowledge Base, which is a database of cleansing rules definition. These DQ functions can be embedded into both FedSQL and DS2 programs. The data quality methods use data quality rules from SAS Quality Knowledge Base to cleanse data.

Figure 6 shows the DQ functions that are supported in this release.

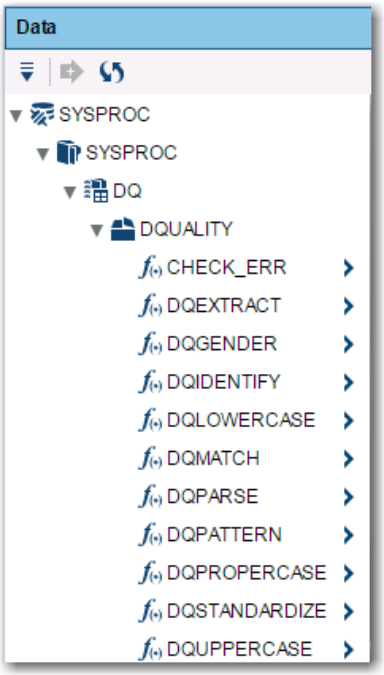

Figure 6. The SYSPROC Data Service in SAS Federation Server 4.2 Hosts the Data Quality Functions

The DQ functions available in this release provide support for the following cleansing methods:

Definition	Method	Description	Example
Case	DQUPPERCASE DQLOWERCASE DQPROPERCASE	Use case definitions to apply uppercase and lowercase lettering using context-sensitive rules.	JOHN SMITH => John Smith
Extraction	DQEXTRACT	Extraction definitions are used to extract specific entities or attributes from a text string.	Blue men's long-sleeved button-down collar denim shirt
Gender	DQGENDER	Use gender definitions to determine the gender of a person from their name or other information.	Jane Smith = F Sam Adams =M
Identification	DQIDENTIFY	Identification definitions determine the type of data that is represented by a text string.	John Smith = Name SAS = Organization
Match	DQMATCH	Use match definitions to generate a matchcode for a text string.	John Smith J. Smith Mr. Jon Smith
Parse	DQPARSE	Use parse definitions to segment a string into several parts.	Mr. Roy G Biv Jr
Pattern	DQPATTERN	Use pattern definitions to return a simple representation of a character pattern based on a text string.	999-999-999
Standardize	DQSTANDARDIZE	Returns a character value after standardizing casing, spacing, and format.	919.6778000 => (919) 677-8000

We'll demonstrate data cleansing example using the scenario from the "Exploring Data Access Control Strategies for Securing and Strengthening Your Data Assets Using SAS® Federation Server" paper. This example uses data from a bank transactions table. Unfortunately, the STATE column has data in a format inconsistent with the requirement (Figure 7).

#	ID	ACCOUNT_NUM	TRANSACTION_TYPE	TRANSACTION_AMOUNT	TELLER_TYPE	STATE	DATE
1	1111	1234567890	DEPOSIT	500.0	ATM	PA	10DEC2015
2	2222	2345678901	DEPOSIT	100.0	MOBILE	VA	21JAN2016
3	3333	3456789012	INQUIRY	1433.45	ATM	CA	02FEB2016
4	4444	1234567890	WITHDRAWAL	-250.0	BANK	MD	19NOV2015
5	5555	9991112222	PAYMENT	1200.5	ONLINE	MA	01JAN2016

Figure 7. Output from Bank Transactional Data. STATE Values Do Not Match the Requirement.

We will create a view stored under the Base SAS data service that will standardize values from the STATE column in accordance with the requirements of SAS Quality Knowledge Base deployed with SAS Federation Server. Use the SYSPROC.DQ.DQUALITY.DQSTANDARDIZE function in the FedSQL view definition to complete this task (Figure 8).

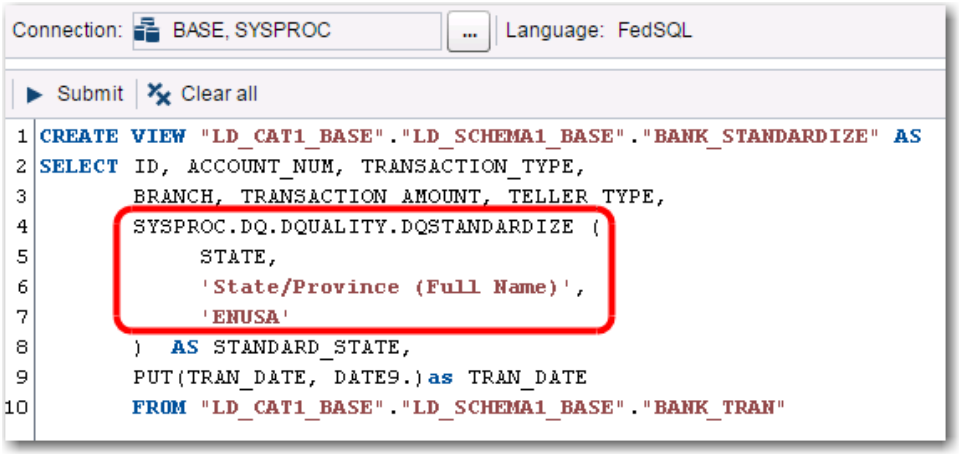

Figure 8. FedSQL Code Example That Creates a View Using the DQSTANDARDIZE Function

Issuing a SELECT statement against this view returns the output in Figure 9. Notice that the value of STATE is spelled out as required.

#	ID	ACCOUNT_NUM	TRANSACTION_TYPE	TRANSACTION_AMOUNT	TELLER_TYPE	STANDARD STATE	DATE
1	1111	1234567890	DEPOSIT	500.0	ATM	Pennsylvania	10DEC2015
2	2222	2345678901	DEPOSIT	100.0	MOBILE	Virginia	21JAN2016
3	3333	3456789012	INQUIRY	1433.45	ATM	California	02FEB2016
4	4444	1234567890	WITHDRAWAL	-250.0	BANK	Maryland	19NOV2015
5	5555	9991112222	PAYMENT	1200.5	ONLINE	Massachusetts	01JAN2016

Figure 9. Output from Bank Transactional Data. STANDARD STATE Values Are Presented as Required.

DQ functions embedded into DS2 programs can be enabled to execute in database. For this capability, data source should reside in a data platform that supports SAS Data Quality Accelerator (in the third maintenance release for SAS 9.4, these platforms are Teradata and Hadoop), and SAS Data Quality Accelerator should be licensed and configured on site.

HOW DO I MASK SENSITIVE INFORMATION?

A common way to limit the exposure of sensitive (personally identifiable) information is to prohibit access to this data by restricting reading from the corresponding source data columns or by designing a secured abstract layer of views that exclude such columns.

There are times though when this sensitive data might provide additional analytical insights. As a result, this data needs to exist in data feeding the analysis and reports. Data substitution algorithms (known as data masking rules) allow you to dynamically substitute sensitive data with encoded values while maintaining critical data characteristics.

The concept of data masking is not new to SAS Federation Server 4.2, but this release provides additional data masking rules to widen the number of dynamic masking application scenarios.

Figure 10 shows the list of functions available in the 4.2 release:

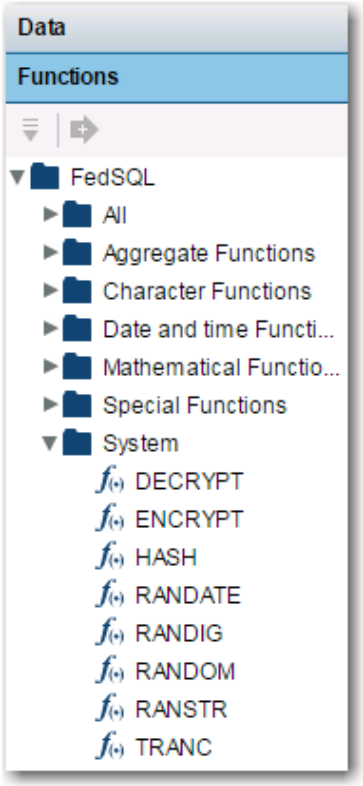

Figure 10. List of Data Masking Functions in SAS Federation Server 4.2

The new masking rules are TRANC and a series of RANDOM-based rules: RANDIG, RANSTR, and RANDATE.

- **The TRANC Rule Type** masks the values by transliterating characters from an input string to characters in an output string. It incorporates logic to map multiple input characters into a single output value to enhance the encoding.
- **The RANDIG Rule Type** masks the numeric values in a column by replacing digits with strings of random digits.
- **The RANSTR Rule Type** masks the values in a column by replacing the values with random strings.
- **The RANDATE Rule Type** masks the values in a date column by replacing them with pseudo-random date values.

To embed the dynamic masking function into a query or a view definition, use the SYSCAT.DM.MASK function with the corresponding rule type parameter.

See Figure 11 for a code example of embedding a RANDATE function in a FedSQL query to mask the DATE column in a bank transaction table.

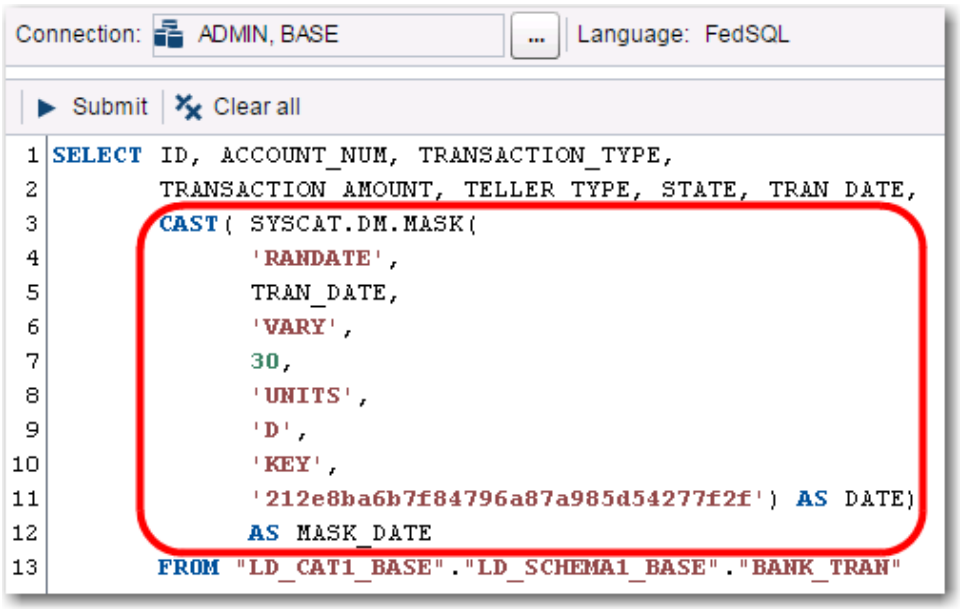

Figure 11. Example of Embedding a RANDATE Masking Function

This code produces the output in Figure 12. Note how values in the TRAN_DATE column are mapped to random data values in the MASK_DATE column. You might also notice that random values are within a 30-days proximity of the original date. This and other controls are built in as parameters for masking functions.

#	ID	ACCOUNT_N...	TRANSACTIO...	TRANSACTIO...	TELLER_TYPE	STATE	TRAN_DATE	MASK_DATE
1	1111	1234567890	DEPOSIT	500.0	ATM	PA	12/10/15	11/29/15
2	2222	2345678901	DEPOSIT	100.0	MOBILE	VA	01/21/16	02/06/16
3	3333	3456789012	INQUIRY	1433.45	ATM	CA	02/02/16	02/18/16
4	4444	1234567890	WITHDRAWAL	-250.0	BANK	MD	11/19/15	11/28/15
5	5555	9991112222	PAYMENT	1200.5	ONLINE	MA	01/01/16	01/28/16

Figure 12. Output from the RANDATE Function Appears in the MASK_DATE Column

CONCLUSION

SAS Federation Server is the solution you want when you have these challenges:

- your enterprise data architecture consists of heterogeneous data platforms that collectively feed into analytical and business analysis applications but joining and managing access to this diverse data by a wide variety of users becomes problematic
- access to your data should be carefully governed
- data requests for federated data should be orchestrated with performance optimization in mind

SAS Federation Server 4.2 will help you design and maintain a logical data warehouse layer that is consistently secured and governed for data access and manipulations across your data environments.

REFERENCES

Craver, Mark and Mike Frost. 2014. “Exploring Data Access Control Strategies for Securing and Strengthening Your Data Assets Using SAS® Federation Server.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary, NC: SAS Institute Inc. Available at https://support.sas.com/resources/papers/proceedings14/SAS394-2014.pdf

SAS Institute Inc. 2016. *SAS® Federation Server 4.2: Administrator’s Guide*. Available at http://support.sas.com/documentation/cdl/en/fedsrvag/68546/PDF/default/fedsrvag.pdf

SAS Institute Inc. 2016. *SAS Federation Server Manager 4.2: User’s Guide*. Available at http://support.sas.com/documentation/cdl/en/fedsrvmgrug/67835/PDF/default/fedsrvmgrug.pdf

SAS Institute Inc. 2016. *SAS 9.4 LIBNAME Engine for SAS Federation Server: User's Guide*. Available at http://support.sas.com/documentation/cdl/en/engfedsrv/67229/PDF/default/engfedsrv.pdf

SAS Institute Inc. 2015. *SAS 9.4 FedSQL Language Reference*. 4th ed. Available at http://support.sas.com/documentation/cdl/en/fedsqlref/67954/PDF/default/fedsqlref.pdf

SAS Institute Inc. 2015. *SAS 9.4 DS2 Language Reference*. 5th ed. Available at http://support.sas.com/documentation/cdl/en/ds2ref/68052/PDF/default/ds2ref.pdf

SAS Institute Inc. 2015. *SAS 9.4 In-Database Products: User's Guide*. 6th ed. Available at http://support.sas.com/documentation/cdl/en/indbug/68442/PDF/default/indbug.pdf

ACKNOWLEDGMENTS

The author’s sincere gratitude goes to these colleagues for their support and knowledge sharing:

Barbara Deaton

Marie Dexter

Brian Hess

Ivor Moan

Jeff Stander

Johnny Starling

Cindy Wang

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Tatyana Petrova
SAS Institute Inc.
919-677-8000
Tatyana.Petrova@sas.com
www.sas.com